Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n2p350

Adjusted R^2 - type measures for beta regression model

By Mahmood, Seyala, Algamal

Published: 14 October 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 02, October 2020, 350-357
DOI: 10.1285/i20705948v13n2p350

Adjusted R^2 - type measures for beta regression model

Shaimaa Waleed Mahmood, Noor Nawzat Seyala, and Zakariya Yahya Algamal*

Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

Published: 14 October 2020

R^2 measure, which named coefficient of determination, is usually used as tools for evaluation the predictive power of the regression models. However, this measure, which is based on deviance for generalized linear models, is sensitive to the small samples. Therefore, it is necessary to adjust R^2 measure according to the number of covariates. Beta regression model has received much attention in several science fields in modeling proportions or rates data. In this paper, several adjusted R^2 measures are proposed in beta regression models. The performance of the proposed measures is evaluated through simulation and real data application. Results demonstrate the superiority of the proposed measures compared to others.

keywords: Deviance, beta regression, coefficient of determination.

1 Introduction

The R^2-type measures, also named explained variation or coefficient of determination gives information about a regression model that as well the information gave by associated P-value and parameter estimates (Heinzl and Mittlböck, 2003). R^2 measure is well-determined and useful tool to assess regression model analysis and becomes more and more familiar in generalized linear models (Martina and Harald, 2001; Mittlboeck and Waldhoer, 2000; Ricci and Martínez, 2008).

In recent years, many papers have transacted with R^2 measures for Poisson regression models. The base advantage has often been in the behavior of many R^2-type measures without considering the number of fitted parameters (W et al., 2014; Waldhoer et al.,

*Corresponding author: zakariya.algamal@uomosul.edu.iq

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1998). Dunn and Smyth (2005) developed a numerical algorithm to assess the densities for Tweedie models and obtaining maximum likelihood estimators.

14 Oc The beta regression model has received much interest in many science fields in rates data or modeling proportions (Algamal, 2019; Uraibi et al., 2017). Ferrari and Cribari-Neto (2004) presented a regression model in which the dependent variable is beta distribution". Recently, several studies of this model are considered (Espinheira et al., 2008; Ferrari and Cribari-Neto, 2004; Cameron and Windmeijer, 1996; Ospina, 2006; Smithson and Verkuilen, 2005).

In this paper, several adjusted R2 measures are proposed in beta regression models. The performance of the proposed measures is evaluated through simulation and real data application. The remainder of this paper is organized as follows, section 2 describe the R^2 measures and their adjustments. Section 3, presents the beta regression model. In Section 4, the R^2 measures and adjustments for beta regression model is explained. Section 5 and 6, presents a real data application and simulation study. The conclusion is covered by Section 7.

2 R^2 measures and Adjustments

The R^2 measures are much used in linear regression model and research has condensed on their application in other generalized linear models (Zheng, 2000). It is well known that W et al. (2014) and Waldhoer et al. (1998) presented the R^2- type measures based on deviances for regression models. The R^2 values are increases and biased as the number of covariates grows, except when the parameter estimates are zero then the R^2 value unchanged. They have properties that make them a highly useful tool in diagnostics and model selection, they take values in $0 \leq R^2 \leq 1$ and they are non-decreasing as regresses are added (Ricci and Martínez, 2008). The R^2 measure based on deviance residuals for regression models defined as

$$R^2_{DEV} = 1 - \frac{D(y,\hat{\mu})}{D(y,\bar{y})} = \frac{\log L(y) - \log L(\hat{\mu})}{\log L(y) - \log L(\bar{y})} \tag{1}$$

where $D(y,\hat{\mu})$ *and* $D(y,\bar{y})$, are the deviance of the full model and the deviance of the null model respectively Mittlböck (2002). The Waldhoer et al. (1998) proposed an adjustment by the degree freedom to the R^2 measure adjustment in linear regression because the inflation of R^2- type measures can be considerable when the number of covariates is major relative to a specific sample size which that defined as

$$R^2_{DEV,df} = 1 - \frac{(n-k-1)^{-1}D(y,\hat{\mu})}{(n-1)^{-1}D(y,\bar{y})} \tag{2}$$

where n *and* k are the sample size and the number of estimated covariates without intercept (Mittlboeck and Waldhoer, 2000). Mittlböck (2002), were proposed two alternative adjustments for R^2 measures based on deviance residuals, defined as

$$R^2_{DEV,adj,1} = 1 - \frac{D(y,\hat{\mu}) + \frac{k}{2}}{D(y,\bar{y})} = 1 - \frac{\log L(y) - \log L(\hat{\mu}) + \frac{k}{2}}{\log L(y) - \log L(\bar{y})} \tag{3}$$

$$R^2_{DEV,adj,2} = 1 - \frac{D(y,\hat{\mu}) + \frac{(k+1)}{2}}{D(y,\bar{y})} = 1 - \frac{\log L(y) - \log L(\hat{\mu}) + \frac{(k+1)}{2}}{\log L(y) - \log L(\bar{y})} \tag{4}$$

Clearly, $R^2_{DEV,adj,2}$ constantly gives values closer to zero than $R^2_{DEV,adj,1}$".

3 Beta regression model

The beta regression model as proposed by Ferrari and Cribari-Neto (2004) and based on the assumption that the response (dependent) variable is beta distribution (Bayer and Cribari-Neto, 2013). "The probability density function for response variable Y can be defined as

$$f(y,\mu,\phi) = \frac{\Gamma\phi}{\Gamma(\mu\phi)\,\Gamma((1-\mu)\phi)}\, y^{\mu\phi-1}(1-y)^{((1-\mu)\phi)-1} \quad 0 < y < 1 \tag{5}$$

where $0 < \mu < 1$ *and* $\phi > 0$, Γ denotes the gamma function, ϕ is a precision parameter. The mean and the variance of Eq.(5) are defined as

$$E(Y) = \mu, \qquad Var(Y) = \frac{\mu(1-\mu)}{1+\phi}$$

For a fixed value ofμ, the higher the value ϕ, the lower the variance of Y (Aktaş and Unlu, 2017). Let regression data $\{(x_j, y_j)\}_{j=1}^n$ each $y_j \sim beta(\mu\phi, (1-\mu)\phi)$, $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$ is an explanatory variable vector, then in beta regression model, the mean is related to the explanatory variables as

$$g(\mu_j) = x_j^T \beta = \eta_j \tag{6}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of unknown regression coefficient. Logit, loglog, probit and Cauchy are the used link function of Eq.(6). The likelihood function can be written as

$$\begin{aligned} &L(\beta,\gamma) = \textstyle\sum_{j=1}^n \ell j(\mu_j,\phi) = \sum_{j=1}^n [\log\Gamma(\phi) - \log\Gamma(\mu_j\phi) - \log\Gamma((1-\mu_j)\phi) \\ &+(\mu_j\phi - 1)\log y_j + \{(1-\mu_j)\phi - 1\}\log(1-y_j)] \end{aligned} \tag{7}$$

where $\mu_j = g^{-1}(\eta_j)$ (Magalhães et al., 2013)".

4 The proposed R^2- type measures for beta regression model

It is well known that Waldhoer et al. (1998) proposed R^2 measures for Poisson regression model, Ferrari and Cribari-Neto (2004) introduced deviance residual which is based on Espinheira et al. (2008); Espinheira and Cribari-Neto (2014)

$$r_j^d = sign(y_j - \hat{\mu}_j)[2\{\ell_j(y_j,\hat{\phi}) - \ell_j(\hat{\mu}_j,\hat{\phi})\}]^{1/2} \tag{8}$$

Here, sign is the sign function. "In this section, we proposed four R^2 measures for beta regression model dependent on R^2 measures rules known. The deviance of beta regression model is proportional to twice the difference between the maximum log likelihood achievable saturated model and that achieved by the model under realization. This is can be written as

$$D(y,\hat{\mu},\hat{\phi}) = \sum_{j=1}^{n}(r_j^d)^2 = \sum_{j=1}^{n} 2\{\ell_j(y_j,\hat{\phi}) - \ell_j(\hat{\mu}_j,\hat{\phi})\} \tag{9}$$

Thus, the R^2 measure based on deviance residuals of beta regression model is

$$R^2_{DEV} = 1 - \frac{D(y,\hat{\mu},\hat{\phi})}{D(y,\bar{\mu},\hat{\phi})} = 1 - \frac{\sum_{j=1}^{n} 2\{\ell_j(y_j,\hat{\phi}) - \ell_j(\hat{\mu}_j,\hat{\phi})\}}{\sum_{j=1}^{n} 2\{\ell_j(y_j,\hat{\phi}) - \ell_j(\bar{\mu},\hat{\phi})\}} \tag{10}$$

such that $D(y,\bar{\mu},\hat{\phi})$ *and* $D(y,\hat{\mu},\hat{\phi})$ are the deviances under the null model and under the full model. In addition, we proposed R^2 measure adjustment by the degree of freedom for this model which be defied as

$$R^2_{DEV,df} = 1 - \frac{(n-k-1)^{-1}D(y,\hat{\mu},\hat{\phi})}{(n-1)^{-1}D(y,\bar{\mu},\hat{\phi})} \tag{11}$$

It has known that n *and* k are the sample size and the number of estimated covariates without intercept. It then we proposed two alternative adjustments for R^2 measures based on deviance residual of this model which can be write as follows

$$R^2_{DEV,adj,1} = 1 - \frac{D(y,\hat{\mu},\hat{\phi}) + \frac{k}{2}}{D(y,\bar{\mu},\hat{\phi})} \tag{12}$$

$$R^2_{DEV,adj,2} = 1 - \frac{D(y,\hat{\mu},\hat{\phi}) + \frac{(k+1)}{2}}{D(y,\bar{\mu},\hat{\phi})} \tag{13}$$

5 Application

In this section, a real data application is presented for finding our proposed R^2- type measures for beta regression model to dataset from body fat, which is include 252 observation for body fat patients on 13 explanatory variables and the response variable. In this dataset as shown in Table 1

Table 1: Features of data body fat

Variables	Features	Variables	Features
y	Percentage of body fat	x_7	Hip circumference
x_1	Age (year)	x_8	Thigh circumference
x_2	Weight (pounds)	x_9	Knee circumference
x_3	Height (inches)	x_{10}	Ankle circumference
x_4	Neck circumference	x_{11}	Extended biceps circumference
x_5	Chest circumference	x_{12}	Forearm circumference
x_6	Abdomen circumference	x_{13}	Wrist circumference

We used four models for the beta regression model to this data. Model 1, includes five explanatory variables. Model 2, includes eight explanatory variables. Model 6, includes tens explanatory variables. Model 4 includes thirteen explanatory variables. The results of the proposed R^2 measures for the four models are summarized in Table 2. It can be seen from Table 2 that the R^2 measures clearly increases when the covariates added to model and all $R^2_{DEV,adj,2}$ values yields lower than $R^2_{DEV,adj,1}$ values.

Table 2: Estimated R^2 measures and their adjustments

Models	k	R^2_{DEV}	R^2_{DEV,d_f}	$R^2_{DEV,adj,1}$	$R^2_{DEV,adj,2}$
Model 1	5	0.5512	0.5421	0.5418	0.5407
Model 2	8	o.7042	0.6944	0.6943	0.6935
Model 3	10	0.7051	0.6929	0.6929	0.6920
Model 4	13	0.7191	0.7037	0.7040	0.7031

6 Simulation Study

In this section, the performance of the four R^2 measures is compared under various conditions for the beta regression model. The sample size is considered with $n = \{50, 100, 200, 300\}$ and precision parameter$\phi = \{20, 50\}$. The response variable was generated from the beta distribution $y_j \sim beta(\mu_j\phi, (1-\mu_j)\phi)$, such that μ_j is generated

according to the logit link function which is defined as

$$\mu_j = \frac{\exp\left(x_j^T \beta\right)}{1 + \exp\left(x_j^T \beta\right)} \tag{14}$$

The variables x_j is generated from the uniform distribution $[0, 1]$. The true parameter vector is $\beta = \{1, -1, 1, 0, 0, 0\}$ and the number of covariates is$k = \{3, 4, 6\}$. The results for the four R^2- type measures are shown in Table 3

Table 3: Simulation results for the proposed measures

ϕ	k	n	R^2_{DEV}	R^2_{DEV,d_f}	$R^2_{DEV,adj,1}$	$R^2_{DEV,adj,2}$
20	3	50	0.8731	0.8648	0.8649	0.8625
		100	0.9261	0.9238	0.9236	0.9228
		200	0.9115	0.9101	0.9100	0.9096
		300	0.8900	0.8889	0.8888	0.8885
	4	50	0.8743	0.8631	0.8638	0.8615
		100	0.9223	0.9191	0.9187	0.9179
		200	0.8883	0.8860	0.8857	0.8851
		300	0.8902	0.8887	0.8886	0.8883
	6	50	0.8727	0.8549	0.8572	0.8550
		100	0.8972	0.8906	0.8906	0.8896
		200	0.8992	0.8961	0.8958	0.8953
		300	0.9010	0.8989	0.8988	0.8984
50	3	50	0.9569	0.9541	0.9543	0.9534
		100	0.9304	0.9282	0.9283	0.9276
		200	0.9547	0.9540	0.9540	0.9537
		300	0.9413	0.9407	0.9407	0.9406
	4	50	0.9418	0.9367	0.9371	0.9359
		100	0.9439	0.9415	0.9416	0.9410
		200	0.9542	0.9532	0.9532	0.9530
		300	0.9559	0.9553	0.9553	0.9552
	6	50	0.9397	0.9313	0.9324	0.9312
		100	0.9453	0.9418	0.9419	0.9414
		200	0.9488	0.9472	0.9472	0.9469
		300	0.9510	0.9500	0.9500	0.9499

From Table 3, it can be seen that the adjusted R^2 measures give less than the R^2 measures unadjusted and always the $R^2_{DEV,adj,2}$ gives values lower than the $R^2_{DEV,adj,1}$. The R^2 measures increases with ϕ.

7 Conclusions

In this work, the problem of adjusting the R^2 measures in beta regression model is investigated. Several adjusting were proposed. Simulation and real data application are carried out. The obtained results prove the dominance of the proposed adjusting R^2 measures against the unadjusted measure.

References

Aktaş, S. and Unlu, H. (2017). Beta regression for the indicator values of well-being index for provinces in turkey. *Engineering Technology and Applied Sciences*, 2(2):101–111.

Algamal, Z. Y. (2019). A particle swarm optimization method for variable selection in beta regression model. *Electronic Journal of Applied Statistical Analysis*, 12:508–519.

Bayer, F. M. and Cribari-Neto, F. (2013). Bartlett corrections in beta regression models. *Journal of Statistical Planning and Inference*, 143(3):531–547.

Cameron, A. and Windmeijer, F. (1996). R2 measures for count data regression models with applications to health care utilization. *Business Econom. Statist.*, 14:209–220.

Dunn, P. and Smyth, G. (2005). Series evaluation of tweedie exponential dispersion model densities. *Statist. Comput.*, 15(4):267–280.

Espinheira, P. L., Ferrari, S. L. P., and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419.

Espinheira, P. L., S. F. and Cribari-Neto, F. (2014). Bootstrap prediction intervals in beta regressions. *Computational Statistics*, 29(5):1263–1277.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Heinzl, H. and Mittlböck, M. (2003). Pseudo r-squared measures for poisson regression models with over- or underdispersion. *Computational Statistics & Data Analysis*, 44(1-2):253–271.

Magalhães, T. M., Botter, D. A., and Sandoval, M. C. (2013). Asymptotic skewness for the beta regression model. *Statistics & Probability Letters*, 83(10):2236–2241.

Martina, M. and Harald, H. (2001). A note on r2 measures for poisson and logistic regression models when both models are applicable. *Journal of Clinical Epidemiology*, 54:99–103.

Mittlböck, M. (2002). Calculating adjusted r2 measures for poisson regression models. *Computer Methods and Programs in Biomedicine*, 68:205–214.

Mittlboeck, M. and Waldhoer, T. (2000). Adjustments for r2-measures for poisson regression models. *Computational Statistics & Data Analysis*, 34:461–472.

Ospina, R., C.-N. F. V. K. L. P. (2006). Improved point and intervalar estimation for a beta regression model. *Computational Statistics & Data Analysis*, 51(2):960–981.

Ricci, L. and Martínez, R. (2008). Adjusted r2-type measures for tweedie models. *Computational Statistics & Data Analysis*, 52(3):1650–1660.

Smithson, M. and Verkuilen, J. (2005). A better lemon-squeezer? maximum likelihood regression with beta distribuited dependent variables. *Psychol Methods*, 11:54–71.

Uraibi, H., Midi, H., and Rana, S. (2017). Selective overview of forward selection in terms of robust correlations. *Communications in Statistics: Simulation and Computation*, 46:5479–5503.

W, Z., R, Z., Y, L., and J, L. (2014). Variable selection for varying dispersion beta regression model journal of applied statistics. *doi:10.1080/02664763.2013.830284*, 41:95–108.

Waldhoer, T., Haidinger, G., and Schober, E. (1998). Comparison of r2 measures for poisson regression by simulation. *Epidemiol. Biostatist*, 3:209–215.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine*, 19:1265–1275.